Correspondences between low-energy modes in enzymes: Dynamics-based alignment of enzymatic functional families

ANDREA ZEN,[1] VINCENZO CARNEVALE,[1] ARTHUR M. LESK,[2] AND CRISTIAN MICHELETTI[1]

[1]International School for Advanced Studies and CNR-INFM Democritos, 34014 Trieste, Italy
[2]Department of Biochemistry and Molecular Biology, The Pennsylvania State University, University Park, Pennsylvania 16802, USA

(RECEIVED December 7, 2007; FINAL REVISION February 13, 2008; ACCEPTED February 13, 2008)

Abstract

Proteins that show similarity in their equilibrium dynamics can be aligned by identifying regions that undergo similar concerted movements. These movements are computed from protein native structures using coarse-grained elastic network models. We show the existence of common large-scale movements in enzymes selected from the main functional and structural classes. Alignment via dynamics does not require prior detection of sequence or structural correspondence. Indeed, a third of the statistically significant dynamics-based alignments involve enzymes that lack substantial global or local structural similarities. The analysis of specific residue–residue correspondences of these structurally dissimilar enzymes in some cases suggests a functional relationship of the detected common dynamic features. Including dynamics-based criteria in protein alignment thus provides a promising avenue for relating and grouping enzymes in terms of dynamic aspects that often, though not always, assist or accompany biological function.

Keywords: structure and function of enzymes; large-scale movements; protein alignment; dynamics-based alignment

Supplemental material: see www.proteinscience.org

Available alignment tools detect similarities among proteins in the first step of the logical cascade: sequence → structure → function. We introduce and apply a general and systematic algorithm to address functional relationships of enzymes by including dynamics in the alignment procedure. We focus on a common although not universal feature of enzymatic function, internal large-scale concerted movements. A large body of evidence links these movements to the structural changes that often accompany protein functions. For example, the displacements involved in allosteric changes in many proteins occur along the collective coordinates corresponding to the low-energy modes of the two stable structures (Delarue and Sanejouand 2002; Falke 2002; Rod et al. 2003; Alexandrov et al. 2005; Ming and Wall 2005; Smith et al. 2005; Zheng et al. 2007).

The collective and large-scale character of these fluctuations has justified their characterization by simplified approaches, typically elastic network models (ENM) (Bahar et al. 1997; Hinsen 1998; Atilgan et al. 2001; Delarue and Sanejouand 2002; Micheletti et al. 2004; Sulkowska et al. 2007). These models rely on a simplified

Reprint requests to: Cristian Micheletti, International School for Advanced Studies and CNR-INFM Democritos, Via Beirut 2-4, 34014 Trieste, Italy; e-mail: michelet@sissa.it; fax: 39-040-3787528.

Abbreviations: ENM, elastic network model; MD, molecular dynamics; RMSD, root mean square distance; RMSIP, root mean square inner product.

Article published online ahead of print. Article and publication date are at http://www.proteinscience.org/cgi/doi/10.1110/ps.073390208.

free-energy function with quadratic dependence on displacements of amino acids from their reference position. This harmonic free-energy approximation would be expected a priori to hold only for very small fluctuations of the protein structure. However, the harmonic approximation to the free energy has proved very valuable for large-scale changes also (Levy et al. 1984; Horiuchi and Go 1991; Brooks et al. 1995; Tirion 1996; Hinsen 1998; Pontiggia et al. 2007). Indeed, linear combinations of the 10 lowest-energy modes predicted by ENMs are generally sufficient to describe most of the conformational fluctuations observed in extensive MD simulations as well as functionally oriented changes between apo and holo forms of enzymes (Delarue and Sanejouand 2002; Falke 2002; Rod et al. 2003; Alexandrov et al. 2005; De los Rios et al. 2005; Ming and Wall 2005; Smith et al. 2005; Zheng et al. 2007).

Here we apply the collective low-energy modes of residues determined by these simplified dynamic models to protein alignment. Unlike in structural alignments, matched residues need show only loose spatial proximity. The spatial tolerance is such that the relative movements in the two enzymes are well defined, yet sufficiently generous to establish correspondences between, for example, different types of secondary structure elements.

This work extends recent studies (Carnevale et al. 2006; Capozzi et al. 2007), in which common features were detected among the low-energy modes of proteolytic enzymes and EF-hand motifs, the structural alignments of which were known. Here we avoid the asymmetric treatment of structural and dynamic features by using a novel optimization scheme that identifies the set of residues that has the highest consistency of large-scale displacements, within tolerant structural correspondence. Combining structural and dynamic criteria on an equal footing appears to be necessary to detect general analogies of the internal motion of biomolecules. A pure dynamic alignment, that is, rewarding the consistency of the low-energy modes' directionality in two sets of residues regardless of their relative spatial relationship would, in fact, not necessarily identify *regions* that undergo analogous dynamic modulations. At the same time, the matching of the ENM-derived low-energy modes is not reducible to establishing correspondences of simple local geometric features of two protein structures. The algorithm, in fact, goes beyond capturing correspondences between the profiles of amino acid mobility, which largely reflect static local structural (density) features (Halle 2002), and promotes the accord of nonlocal correlations of amino acid displacements in thermal equilibrium. In view of the collective, nonlocal, nature of the ENM-derived equilibrium fluctuations exploited by the algorithm it appears justified to term the alignment as *dynamics-based*.

The alignment procedure is applied to all pairs from a set of 76 enzymes which represent the main functional families with minimal structural redundancy. The alignment score of ~30 enzyme pairs was found to be outstanding by standard criteria of statistical significance. Two-thirds of such alignments reflect global or partial correspondences in the fold architecture. Notably, the remaining third involve proteins with only loose analogies of secondary and tertiary structural elements but with precisely matching large-scale dynamics. Even for structurally dissimilar pairs of enzymes the dynamics-based alignment can induce a remarkable spatial superposition of functionally relevant regions. This suggests a biological rationale underlying specific common concerted movements. Further development of tools capable of detecting such dynamic correspondences is expected to provide novel elements and perspectives to address the relationships among sequence, structure, and function of enzymes.

Results and Discussion

We searched for common large-scale movements in pairs of enzymes chosen as representative of different functional classes according to the Enzyme Commission (EC) (PDB codes were taken from http://www.ebi.ac.uk/thornton-srv/databases/CSA/). From each of the six main EC classes, we selected one representative of each group of molecules sharing the same structural class, architecture, and topology, as defined by the CATH classification of protein structural patterns (Pearl et al. 2005). The resulting working set contained 76 proteins providing a comprehensive and heterogeneous coverage of enzymatic functions and structures (the list of representatives is provided in the Supplemental material). As the removal of structural redundancy was carried out for each EC class separately, representatives of different functional families can have the same topology, according to the CATH classification. This degeneracy, which affects only 47 of the 2850 distinct pairings of the 76 representatives, was retained as its removal would have led to an uneven representation of the distinct EC families.

Each pair of enzymes was considered for the low-energy-modes alignment procedure which is represented schematically in Figure 1 (for details, see Materials and Methods). As in other contexts (Holm and Sander 1996; Altschul et al. 1997; Notredame et al. 2000; Chenna et al. 2003; Lesk 2004; Shatsky et al. 2004a,b; Konagurthu et al. 2006) any alignment establishes a one-to-one pairing among ''marked'' amino acids in the two proteins, plus gaps indicating insertions and deletions. For each pair of representatives we considered thousands of alignments involving increasing numbers of residues, $n = 75$, 100, 125, . . ., up to the maximum length dictated by the

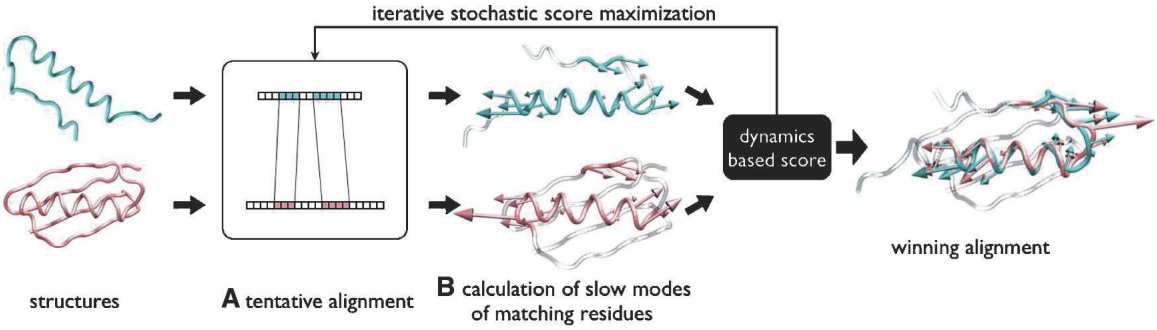

Figure 1. Schematic diagram of the dynamics-based alignment. For each pair of enzyme structures (*left*) thousands of tentative alignments, that is, one-to-one correspondences of amino acids, are considered. For each alignment (*A*) the aligned residues are superposed, and (*B*) for each structure, the low-energy modes of the aligned residues are calculated within the elastic network model. A numerical score, to evaluate the quality of each alignment considered, measures the consistency between the structural alignment and the dynamics reflected in the low-lying normal modes. That is, a particular residue–residue correspondence in the alignment contributes favorably to the score if the residues are *both* well-superposed spatially, and show similar patterns of displacement in the low-lying modes. The optimal alignment is identified by maximizing this score through a stochastic optimization loop. Images of protein structures and low-energy modes were produced with the VMD graphical package (Humphrey et al. 1996).

shorter chain. The best alignment was found by maximizing a scoring function that rewards correspondences between residues with *both* good structural superposition *and* good accord of the low-energy modes (see Materials and Methods). The low-energy modes of the marked residues are calculated within an elastic network approach. This allows a transparent modeling of the potential of the mean force governing the effective interactions of the marked residues that accounts both for direct ''contact'' interactions and for indirect ones mediated by the nonaligned residues.

For each enzyme pair, a stochastic optimization technique produced the best-scoring alignment. The resulting scores are graphically represented in Figure 2 in which the two matrices of panels A and B differ only in the way the entries are ordered. In A, rows and columns appear in order of EC code; in B, in order of CATH code. These two alternative groupings allow an intuitive perception of how functional and structural analogies are reflected by the alignment score. The qualitative appearance of the two plots is markedly different. The minimally redundant coverage of the different EC families produces a fairly uniform scatter of good scores across various EC groups (Fig. 2A). It is nevertheless interesting to notice the presence of light bands corresponding to EC groups that are poorly alignable in general. The most notable of such groups comprises hydrolases acting on acid anhydrides (principal EC codes: 3.6). By contrast, the uneven

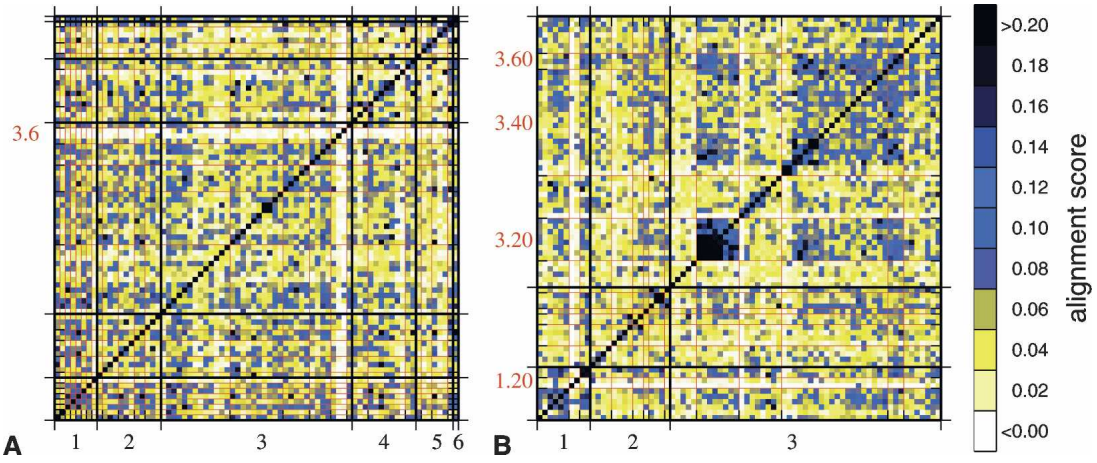

Figure 2. Matrices reporting the dynamics-based score for all aligned enzyme pairs. Good (poor) alignment scores are shown with dark blue (white) color. (*A*) Enzymes are ordered in each axis according to EC codes, black and red lines delimit enzymes with the same first and first two EC numbers, respectively. (*B*) Enzymes are ordered according to CATH codes. Black (red) lines separate different classes (architectures).

representation of different structural classes, architectures, and topologies in the data sets leads to a manifest inhomogeneous character of the matrix ordered by the structure of Figure 2B. In particular, the class with the largest proportion of good scores is the α-β one (class 3), which is also the most populated class in the set. Not all its architectural subgroups, however, display the same degree of ''alignability.'' Both in absolute and relative terms, the most prominent architecture is the α-β barrel (principal CATH numbers: 3.20). Also worth noting is that good alignment scores are attained for several *inter*architecture alignments; a few of such cases will be discussed later. Finally, a noticeable case of overall poor alignability is the set of the mainly α/up-down bundle (principal CATH numbers: 1.20), which shows correspondences only with enzymes belonging to the same structural group.

The extent to which the various degrees of structural relatedness impact on the dynamic correspondences is summarized in Figure 3A. The histogram portrays the distribution of optimized scores for all enzyme pairs and also pairs having the same class, architecture, and topology. It is noted (see inset) that the very few pairings (47 entries) of enzymes with the same topology tend to have alignment scores distinctly better than typical enzyme pairs. On the other hand no such pronounced deviation from the average behavior is observed for pairs with the same structural class or even architecture (that is, the two highest levels of the hierarchical structural classification in CATH).

Similarities in dynamics between structurally related enzymes is expected. We therefore wish to focus particularly on the alignments that are highest ranking according to the dynamics-based score. The distribution of their scores is shown in Figure 3B. In this figure, alignments among enzymes with the same topology (the first three CATH numbers) and the same topology plus homology (the entire CATH code) have been highlighted. Among the top ~20 alignments are six pairs sharing the full CATH code (the total number of such homologous pairs in the set is eight). This confirms the intuitive expectation that significant sequence and structural similarities likely result in pronounced dynamic similarities (Keskin et al. 2000).

However, it is important to note that in Figure 3B, besides these expected good correspondences, a fraction of the alignments approaching the tail pertain to pairs that differ at the level of class or architecture. These cases are of particular interest as they would not be singled out by criteria based solely on the CATH structural classification. A selection of these alignments, as well as other structurally induced ones, will be discussed in the following sections.

Our considerations will be restricted to the alignments that are statistically significant. The significance analysis was performed by comparing the observed scores of

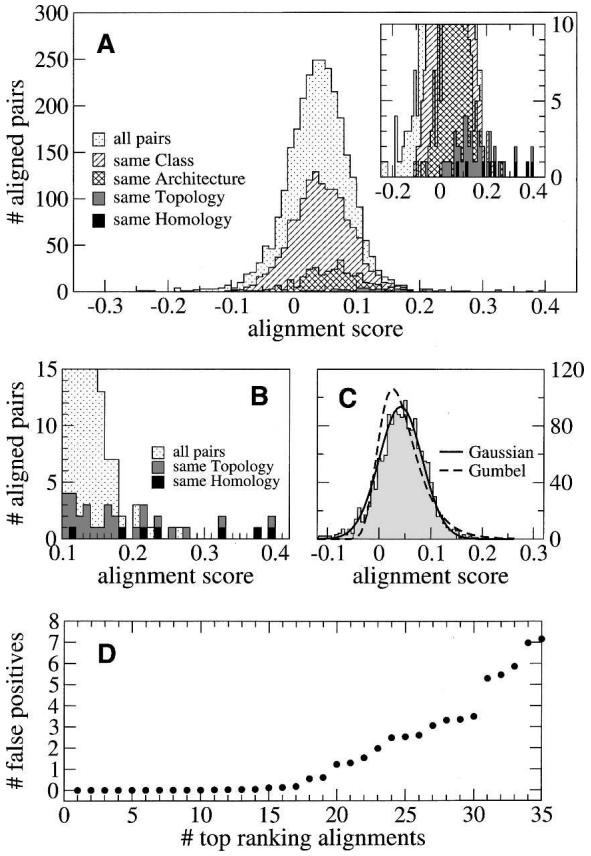

Figure 3. (*A*) Distribution of the alignment score calculated over all 2850 enzyme pairs (the *inset* presents an enlargement of the histogram highlighting pairs with the same topology). The contribution of pairs with the same structural class, architecture, topology, and homologous superfamily are also shown. (*B*) Tail of the distribution associated with the highest alignment scores. Pairs that have the same topology and homologous superfamily are highlighted. (*C*) Filled histogram: distribution of alignment scores over the 56 representatives (highlighted in Table 1) of different topology. The continuous and dashed lines are the best fits to the histogram using Gaussian and Gumbel distributions. Parameters (mean and spread) of the best-fitting Gaussian: $\mu = 0.041413$, $\sigma = 0.041493$. (*D*) Number of nonstatistically significant (false positive) alignments expected to arise within the top-ranking alignments.

Figure 3A with a reference distribution of scores recorded over a set of enzymes that are not expected a priori to lead to a sizeable number of meaningful alignments. In light of previous considerations, this reference set was assembled by selecting one representative, the longest, for each of the 56 different topologies in our data set. The resulting histogram of the 1540 alignment scores was compared against standard statistical distributions arising in alignment contexts (Levitt and Gerstein 1998; Taylor 2006) including the extreme value (Gumbel) and the Gaussian distributions, see Figure 3C. Assuming a Poissonian uncertainty of the height of the histogram, the χ^2 associated to the Gumbel distribution is 3.7, while that of the Gaussian distribution is 1.1. As is visible in Figure

3C, the Gaussian distribution appears to provide a good fit to the data set within three standard deviations to the left and right of the mean value. The latter distribution was consequently taken a posteriori as providing a viable statistical description of the observed measurements and was accordingly used to measure the statistical significance of the top-scoring dynamics-based alignments.

In particular, from the *p*-value analysis we estimated the number of nonsignificant entries (false positives) expected among the top alignments (Levitt and Gerstein 1998; Storey and Tibshirani 2003). The associated curve, shown in Figure 3D, indicates that within the top 26 alignments, fewer than 10% are expected to be false positives. This threshold provides an acceptable balance between the number of entries declared significant (26) and the fraction that is deemed reliable (>90%). All further considerations will therefore be limited to the pairings in the top 26 alignments, which are reported in Table 1.

Within this set, the number of pairings that can be ascribed to overall similarities of the global fold topology is 16, including six homologous cases. A more refined and quantitative study of the level of subtler structural correspondence in the set was carried out with DALI (Holm and Sander 1996; Holm and Park 2000), a powerful structural alignment tool that detects *partial* similarities based on the similarity between two proteins of interresidue distance matrices. For a consistent comparison with our results, the statistical confidence threshold on the DALI results (Sierk and Pearson 2004) was also set to 90% (leading to 18 significant DALI pairings). It was found that 14 of our top 26 alignments had significant DALI scores. These included 12 pairs with the same topology (including all of the six homologous pairs). Of the 10 pairings with different topology selected by our method only two turned out to have significant partial alignments according to DALI. These alignments were between proteins 2dhn-2g64 and 1dy4-2ayh. Importantly, within the 18 statistically significant DALI pairings these two alignments were the only ones involving different CATH topology. Consequently, the remaining eight of the 26 (i.e., ~30%) dynamics-based alignments deemed significant involved pairings between enzymes whose structural relatedness is not easily detectable at the same level of statistical significance.

A selected number of significant alignments, exemplifying the sophisticated interplay of structural and

Table 1. *List of the top 26 dynamics-based alignments*

Rank	PDB1	EC	CATH	Length	PDB2	EC	CATH	Length	n	RMSIP	RMSD (Å)
1	1ajz	2	3.20.20.20	282	1gqn	4	3.20.20.70	252	150	0.8525	5.103
2	1cqh	4	3.40.30.10	105	1mek	5	3.40.30.10	120	75	0.8735	2.773
3	1yb7	4	3.40.50.1820	256	2had	3	3.40.50.1820	310	200	0.8090	4.552
4	1gqn	4	3.20.20.70	252	1nsj	5	3.20.20.70	205	150	0.8051	4.303
5	1ajz	2	3.20.20.20	282	1nsj	5	3.20.20.70	205	150	0.7864	4.067
6	1b1y	3	3.20.20.80	500	1gqn	4	3.20.20.70	252	175	0.7953	5.061
7	2dhn	4	3.30.1130.10	121	2g64	4	3.30.479.10	140	75	0.8261	3.014
8	1b1y	3	3.20.20.80	500	1nsj	5	3.20.20.70	205	125	0.7534	5.988
9	1k03	1	3.20.20.70	399	1nsj	5	3.20.20.70	205	100	0.7777	4.178
10	1id8	5	3.40.50.280	137	1yb7	4	3.40.50.1820	256	75	0.7964	3.606
11	1dbs	6	3.40.50.300	224	1nsj	5	3.20.20.70	205	100	0.7846	5.704
12	1bhe	3	2.160.20.10	376	1vbl	4	2.160.20.10	416	200	0.7167	10.13
13	1ajz	2	3.20.20.20	282	1k03	1	3.20.20.70	399	150	0.7319	7.226
14	1ajz	2	3.20.20.20	282	1b1y	3	3.20.20.80	500	125	0.7785	6.073
15	1gqn	4	3.20.20.70	252	2plc	4	3.20.20.190	274	125	0.7646	6.771
16	1dy4	3	2.70.100.10	434	2ayh	3	2.60.120.200	214	75	0.7493	5.251
17	1ako	3	3.60.10.10	268	1d7o	1	3.40.50.720	297	175	0.7006	8.442
18	1gqn	4	3.20.20.70	252	1k03	1	3.20.20.70	399	200	0.7196	6.302
19	1v3w	4	2.160.10.10	173	1xm0	1	2.170.150.20	147	75	0.6811	11.46
20	1v3w	4	2.160.10.10	173	2dhn	4	3.30.1130.10	121	75	0.6909	11.90
21	1ajz	2	3.20.20.20	282	1dbs	6	3.40.50.300	224	100	0.7869	7.271
22	1id8	5	3.40.50.280	137	2had	3	3.40.50.1820	310	100	0.6823	5.698
23	2f47	3	1.10.530.40	175	4tms	2	3.30.572.10	316	100	0.7159	10.25
24	1gqn	4	3.20.20.70	252	1h17	2	3.20.70.20	754	125	0.7424	9.110
25	1ajz	2	3.20.20.20	282	2plc	4	3.20.20.190	274	100	0.7256	6.446
26	1ako	3	3.60.10.10	268	1avp	3	3.40.395.10	199	75	0.7624	5.990

The first column is the rank of the alignment. Columns 2–9 report the PDB code, principal EC number, CATH code, and length of the aligned proteins. The last three columns provide details of optimal alignment, namely, the number of aligned residues, *n*, the RMSIP of the top 10 modes, and the structural RMSD.

dynamic features, are shown in Figure 4: hydroxynitrile lyase-haloalkane dehalogenases (200 aligned residues) in panel A; human thioredoxin-disulfide isomerase (75 aligned residues) in panel B; dethiobiotin synthetase-phosphoribosyl anthranilate isomerase (100 aligned residues) in panel C; exonuclease III-enoyl reductase (175 aligned residues) in panel D; cellobiohydrolase I-endo-1,3–1,4-β-D-glucan 4-glucanohydrolase (75 aligned residues) in panel E; and exonuclease III-human adenovirus proteinase (75 aligned residues) in panel F. The first two pairs are examples of alignments between enzymes with different functions (first EC number) but similar fold (i.e., same CATH code) while the opposite is true for examples E and F. Cases C and D are, instead, examples of

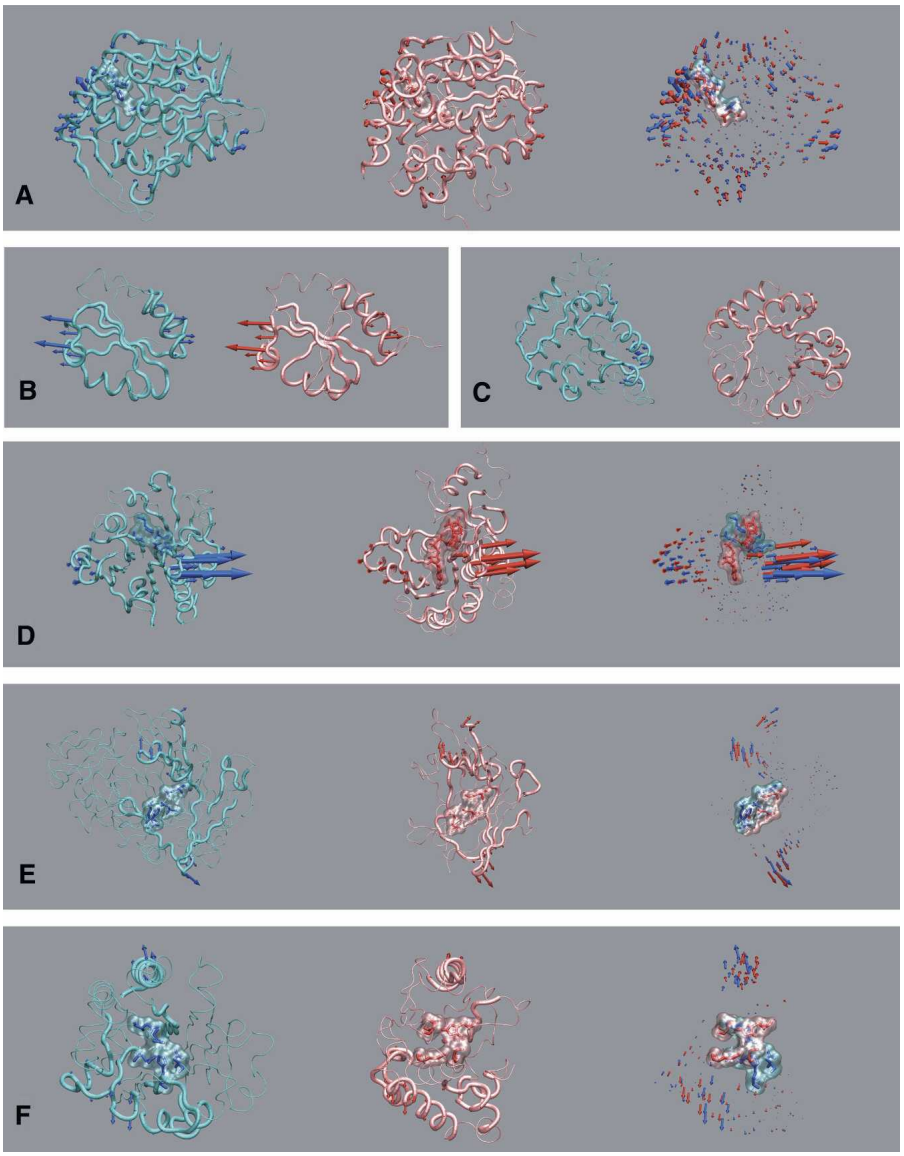

Figure 4. Dynamics-based alignments. Structural-dynamic properties of selected alignments are graphically summarized by rendering, in separate subpanels, in blue the first listed protein and in red the second. Nonaligned (aligned) regions are represented as thick (thin) tubes. Arrows are used to indicate the directionality and magnitude of the distortions entailed by the most consistent dynamic space (see last paragraph of Materials and Methods). (*A*) The portrayed pairs are hydroxynitrile lyase (1yb7) and haloalkane dehalogenase (2had). Additional modes of this pair are shown in the Supplemental material. (*B*) Human thioredoxin (1cqh) and disulfide isomerase (1mek). (*C*) Dethiobiotin synthetase (1dbs) and phosphoribosyl anthranilate isomerase (1nsj). (*D*) Exonuclease III (1ako) and enoyl reductase (1d7o). (*E*) Cellobiohydrolase I (1dy4) and the endo-1,3–1,4-β-D-glucan 4-glucanohydrolase (2ayh). (*F*) Exonuclease III (1ako), and human adenovirus proteinase (1avp). For cases (*A,D,E,F*), the location of the catalytic residues or bound substrates (see text) are highlighted as van der Waals surfaces. The third (*rightmost*) subpanel of these cases (*A,D,E,F*) presents the superposition of the aligned regions.

alignments between enzymes that differ in both function and fold.

One of the enzyme pairs with the highest structural-dynamic correspondence involves hydroxynitrile lyase (PDB: 1yb7, length 256, EC: 4.1.2.39, CATH: 3.40.50.1820) and haloalkane dehalogenase (PDB: 2had, length 310, EC: 3.8.1.5, CATH: 3.40.50.1820). These differ in EC class but have the same first four CATH codes. Their best alignment, which spans 200 residues, covers a substantial fraction of both enzymes. Figure 4A summarizes the results graphically. For clarity, the aligned regions and associated low-energy modes are shown separately for the two enzymes. Given the impossibility of conveying graphically the dynamics covered by the 10 lowest-energy modes, we have reported only the maximally consistent subspace in the two sets of modes (see Materials and Methods).

The RMSD over the 200 aligned residues is 4.5 Å which compares well with the purely structural DALI alignment of the same proteins: RMSD = 3.0 Å over 226 residues. Indeed, unlike other cases discussed in the following, this optimal alignment is also very good from a purely structural point of view. The quality of the overall consistency of the low-energy modes is also striking, as it possesses an RMSIP of 0.81, which exceeds the reference values that typically denote good consistency of molecular dynamics trajectories of the *same* protein (Amadei et al. 1999).

Another high-ranking alignment for both the dynamics-based procedure and the purely structural one is the pair of enzymes human thioredoxin (PDB: 1cqh, length 105, EC: 4.2.99.18, CATH: 3.40.30.10) and disulfide isomerase (PDB: 1mek, length 120, EC: 5.3.4.1, CATH: 3.40.30.10) where as many as 75 residues correspond, with an RMSD as low as 2.8 Å and RMSIP again exceeding 0.87. Panel B of Figure 4 shows the high quality of the accord between structure and dynamics. (Particular enzymes with this fold can show both functions.)

Over a third of the reliable alignments involve pairs that have dissimilar structural organization. Two notable examples appear in panels C and D of Figure 4; for the pairs dethiobiotin synthetase (PDB: 1dbs, length 224, EC: 6.3.3.3, CATH: 3.40.50.300) and phosphoribosyl anthranilate isomerase (PDB: 1nsj, length 205, EC: 5.3.1.24, CATH: 3.20.20.70) in panel C; and exonuclease III (PDB: 1ako, length 268, EC: 3.1.11.2, CATH: 3.60.10.10) and enoyl reductase (PDB: 1d7o, length 297, EC: 1.3.1.9, CATH: 3.40.50.720) in panel D. Even though no strong global structural correspondences can be established between these pairs, there is a discernible consistency of the aligned regions. For the 100 aligned residues of the pair in Figure 4C and 175 aligned residues in Figure 4D, the RMSD values are 5.7 Å and 8.4 Å, respectively. The structural tolerance of this dynamics-based alignment is such that even elements with different secondary organization can be put in structural correspondence (e.g., loops and helices). In these two cases also, low-energy modes are in very good agreement (RMSIP equal to 0.78 and 0.70, respectively) and outline a consistent movement of fairly large compact regions in the enzyme pairs.

Another interesting observation concerns the spatial proximity of the catalytic sites induced by dynamics-based alignments. Bartlett et al. (2003) have shown that evolutionarily distantly related enzyme pairs that catalyze different reactions on similar structural scaffolds, retain the location of the active site and of functional structural elements, suggesting that evolution acts by changing roles and identities of residues at certain positions rather than recruiting new positions. Those observations raise the possibility that, besides local structural patterns, also the dynamic modulations of the active site region have played a role in such conservation (Maguid et al. 2006; Sacquin-Mora et al. 2007). These observations prompted us to investigate whether any of the dynamics-based pairings induce correspondences of features related to catalysis or substrate binding.

Indeed, in our analysis we found that several high-ranking alignments bring active site residues into proximity. The rightmost panel in Figure 4A shows the superposition of the 200 aligned residues hydroxynitrile lyase (PDB: 1yb7) and haloalkane dehalogenase (PDB: 2had), which, being evolutionarily related, are characterized by the same *four* CATH numbers: 3.40.50.1820, and belong to hydrolases (EC class 3) and lyases (EC class 4), respectively. Despite the different biological functions of the two enzymes, the positions of their catalytic residues are almost coincidental. In particular HIS235, ASP207, and SER80 of hydroxynitrile lyase are equivalent to HIS289, ASP260, and ASP124 of haloalkane dehalogenase, respectively.

The alignment between the cellobiohydrolase I (PDB: 1dy4) and the endo-1,3–1,4-β-D-glucan 4-glucanohydrolase (PDB: 2ayh), is also noteworthy as they differ at the CATH architecture level, though they share the same fold according to SCOP (Murzin et al. 1995). The enzymes, which are both glycosylases (EC code: 3.2.1), have analogous catalytic residues (Porter et al. 2004): GLU212, HIS228, ASP214, and GLU217 for the first enzyme; and GLU105, ASP107, and GLU109 for the second one. Despite the fact that only 20% of the larger enzyme is involved in the alignment, it is interesting to observe a remarkable space proximity of the two GLU-ASP-GLU triads (Fig. 4E) which, in both cases, are located in an antiparallel β-sheet. Further aspects of this alignment deserve comment. For 1dy4 the active site is found in a cleft delimited by loops and which can accommodate the 1-(isopropylamino)-3-(1-naphthyloxy)-2-propanol ligand (see leftmost panel in Fig. 4E). Also

for 2ayh the active site is surrounded by loops that form a groove that can arguably accommodate the corresponding ligand (see central panel of Fig. 4E). The dynamics-based alignment has singled out a correspondence between the loops delimiting the binding clefts (residues 369–379 and 185–195, respectively, for 1dy4 and 2ayh) and the directions of the matching low-energy modes are intuitively consistent with the opening/closing mechanism related to substrate binding in both enzymes (Divne et al. 1998).

We now turn to specific enzyme pairing whose global/partial structural correspondences are not easily detectable, as indicated by the much higher, and more significant, dynamics-based ranking compared to the one found by purely structural criteria (Holm and Park 2000). Two of these pairings involve the exonuclease III (PDB: 1ako) which is aligned both with the enoyl reductase (PDB: 1d7o), and with the human adenovirus proteinase (PDB: 1avp). As in previous cases, the dynamics-based alignment induces a good superposition of the functionally relevant regions of the exonuclease III and the enoyl reductase. As shown in Figure 4D, in fact, the active site of 1ako is well superimposed with the ligands bound by 1d7o and in both cases the corresponding low-energy modes develop an outward/inward concerted movement in the surroundings of these regions. This relationship is plausible, given the chemical similarity of the ligands that these proteins bind (Mol et al. 1995; Roujeinikova et al. 1999; Pidugu et al. 2004; Stockwell and Thornton 2006).

A close relatedness of the nature of the ligands is also found for the pairing of exonuclease III and human adenovirus proteinase (Fig. 4F). Both enzymes, in fact, bind DNA (in double- and single-stranded forms, respectively). The possibility of establishing a dynamics-based connection between them is particularly interesting as they are not evolutionarily related and are characterized by two different architectures, a four-layer sandwich (CATH: 3.60.10.10) for 1ako and a three-layer ($\alpha\beta\alpha$) sandwich (CATH: 3.40.395.10), for 1avp. Despite these features, the active site of the enzymes is well-superimposed after the dynamics-based alignment. Notably, the aligned region comprises segments of residues that have been previously suggested to be involved in the binding of DNA (Mol et al. 1995; Gupta et al. 2004).

Finally, among the alignments involving two structurally unrelated enzymes, we mention the case of dihydropteroate synthetase (PDB: 1ajz) and dethiobiotin synthetase (PDB: 1dbs). Despite the differences in architecture, α-β barrel (CATH: 3.20.20.20) for 1ajz and a three-layer ($\alpha\beta\alpha$) sandwich (CATH: 3.40.50.300) for 1dbs, and of the catalyzed reactions, the catalytic residues are found in good correspondence and one-to-one pairings can be established between the three catalytic residues of 1dbs and three of the four catalytic amino acids of 1aj7 (Yang et al. 1997; Porter et al. 2004). The C_α distances of such pairings range from 5.4 Å to 7.5 Å.

The specific cases discussed so far provide concrete illustrations of the biological implications of the dynamics-based alignment. They suggest particularly that functional correspondences in protein may be revealed on the basis of similarity in dynamics, thereby complementing available powerful strategies based on similarity at the level of sequence or at the level of structure. It is, in fact, well known that non-homologous enzymes with similar mechanisms can share the spatial configuration of active site catalytic residues. On the basis of this observation it is possible to detect proteins with related functions by identifying similar configurations of catalytic residues. The alignment scheme considered here is motivated by the fact that some enzymes undergo conformational changes as an integral part of their function. This observation has been applied in a spirit analogous to the structure-based inference mentioned above. In particular it is considered that proteins with similar mechanisms might share not only similar configurations of catalytic residues, but also similarities in dynamics, and that these similarities might be detectable computationally. The specific cases discussed here suggest that proteins can show convergent evolution to shared dynamics related to function.

Conclusion

From the comparative analysis of large-scale movements in representatives of different functional categories of enzymes, approximately 30 outstanding alignments are identified using established criteria for statistical significance. Detailed analysis of the results indicates that good dynamic similarities in enzyme pairs can arise even in the absence of strict correspondence of structure or sequence [pairwise sequence alignments (Chenna et al. 2003) among members of the set yield 12% $\pm$ 2% sequence identity on average]. Indeed, one-third of the outstanding pairings involve enzymes with different structural organization at the global or partial fold level.

Strikingly, it is found that, even in the absence of easily detectable structural correspondences, dynamics-based alignment can establish spatial relationships among regions involved in catalysis or substrate binding. In addition, the common dynamic features are oriented toward the structural rearrangements that arguably accompany the enzymatic functionality. This implies that a biological, function-related rationale underlies several of the outstanding alignments (though this is not necessarily true for *all* alignments, as large-scale movements are not expected to be involved in function for every enzyme).

These facts suggest that dynamics-based criteria can be profitably introduced in protein alignment contexts to expose functionally related correspondences that would not be capturable, at the same level of significance, using purely sequence- or structure-based criteria. As a complement to these established techniques, further developments of dynamics-based approaches can contribute novel elements for exploring relationships between sequence, structure, and function of enzymes.

In this respect the results of the present investigation along with previous studies of dynamic relatedness within specific enzymatic families suggest that tools capable of exposing dynamics-based correspondences may provide a general quantitative and natural framework to group proteins according to their large-scale movements (Carnevale et al. 2006; Capozzi et al. 2007). The robustness of the resulting dynamics-based grouping can, in principle, be verified by independent means. From a computational perspective, the consistency of the dynamics of the aligned regions can be verified by comparing the principal components of their covariance matrices (Garcia 1992), or other measures of correlation (Lange and Grubmueller 2006) calculated in extensive MD simulations of the proteins of interest. A conceptually analogous route can also be envisaged experimentally. Measurements of dynamic order parameters in single-point mutants of a reference enzyme can be used to construct a matrix of pairwise dynamic correlations (Mayer et al. 2003).

Furthermore, by cross-referencing results of purely structural and dynamics-based alignment it might be possible to address if, and to what extent, structural and dynamically related functional features have been subjected to different selective pressure. Two extreme scenarios may, in fact, be envisaged behind function-related dynamic correspondences between structurally diverse enzymes. On one hand common large-scale dynamics may reflect features present in ancestral proteins/enzymes preserved during evolution, or they might reflect features selected by the necessity of well-defined movements for biological function (requiring only very general relationships between sequences and structures). Analogous questions have arisen about protein folds: It appears that both convergence and conservation have resulted in the limited number of available folds (Chothia 1992; Denton and Marshall 2001; Lupas et al. 2001; Andreeva and Murzin 2006; Rose et al. 2006).

It would therefore be most interesting to address these issues connected to the evolutionary convergence/conservation of functionally oriented motions, for specific enzymatic families that have been the subject of thorough investigation from an evolutionary perspective (Lesk and Fordham 1996; Xu et al. 1999; Scheeff and Bourne 2005).

Materials and Methods

Data set selection

The enzymes considered here were selected exploiting the hierarchical classification provided by the EC (Enzyme Commission) database. The EC functional annotation provides a transparent, though qualitative, criterion for defining an enzymatic functional distance which was used in the analysis to investigate the existence of correlations between functional and dynamics-based pairwise similarities. The reference data set was constructed by uniformly covering each of the six EC classes: oxidoreductases, transferases, hydrolases, lyases, isomerases, and ligases. The entire EC database was filtered to remove overall structural redundancies within each class. Only single-chain and single domain enzymes (with complete structural information) were treated, because these are the subjects of the CATH classification (indeed, as shown in the Supplemental material, dynamics-based alignment can be performed also with multi-domain proteins). After applying the selection filter, within each class we kept only one representative per topology (first three CATH numbers), by default the longest enzyme. Very few selected enzymes had highly mobile residues at exposed termini which were omitted from the structural description. The resulting set consisted of 76 enzymes with the following functional distribution: oxidoreductases (eight), transferases (12), hydrolases (36), lyases (12), isomerases (seven), and ligases (one). The total structural variability contains 56 different topologies, representing three CATH structural classes and 15 architectures.

Stochastic exploration of dynamics-based alignments

Statistically significant consistencies of large-scale movements were sought for each of the 2850 distinct pairings of the 76 representative enzymes. The degree of consistency of any pairing was established by means of a single scoring parameter, s, measuring the accord of the spatial position and concerted movements of residues in pairwise correspondence on the two enzymes. Given the differences in length and structural organization it is neither feasible nor meaningful to seek a one-to-one correspondence of *all* amino acids in two proteins. We therefore relied on a stochastic exploration of partial, yet statistically significant, correspondences between subsets of n residues between the two proteins. The stochastic search of putative alignments of fixed length n residues is performed by partitioning the n residues in blocks of at least 10 residues for each protein. The block assignment is done independently for each protein; as a result the number of blocks and their lengths are generally different for the two proteins. The residues taking part in the blocks are numbered sequentially from the N to the C terminus. The alignment is defined as the pairwise correspondence in marked residues with the same index. Starting from an initial block assignment in the two proteins we modify its block sequence by merging/splitting or shifting the blocks. Each trial alignment is accepted/rejected with the standard Metropolis criterion (within a replica-exchange scheme) to promote the maximum score, s_n. The optimal alignment is finally found by maximizing s_n over the explored values of $n = 75,100,125,\cdots$ (the largest value of n being determined by the shortest of the two proteins). The maximized value of s_n is taken as the final alignment score.

Low-energy modes of aligned residues

For each tentative alignment it is necessary to identify the lowest-energy modes of the two proteins of the ''marked''

amino acids only, and to compare them in a common Cartesian reference frame. We shall, accordingly, assume that one of the two proteins has been roto-translated to minimize the root mean square distance between the matching residues. After the optimal superposition, a model free energy is introduced to characterize the thermal equilibrium fluctuations of the marked residues. To this purpose we adopted the well-established elastic network approach (Bahar et al. 1997; Hinsen 1998; Atilgan et al. 2001; Delarue and Sanejouand 2002; Micheletti et al. 2004). The model energy is, in fact, constructed by introducing a chain connectivity term plus harmonic interactions between pairs of spatially close C_α's which disfavor changes of C_α's separation from the native value. For small near-native fluctuations, the resulting energy F penalizes the displacement $\delta\vec{x}_k$ of the kth C_α from its reference position in a quadratic manner,

$$F = \frac{1}{2}\sum_{i,j}\delta\vec{x}_i M_{ij}\delta\vec{x}_j \tag{1}$$

The matrix M is symmetric and its entries are calculated according to the prescription of the β-Gaussian network model of Micheletti et al. (2004) which also allows inclusion of the effective interactions among amino acid side chains (whose positions are derived from the C_α centroids from a deterministic geometric construction). Straightforward, the harmonicity of F lends to the characterization of a protein's fluctuations and of the lowest-energy modes in terms of the eigenvalues and eigenvectors of M. Typically, the calculated modes are in very good accord with the essential dynamic spaces obtained from extensive atomistic molecular dynamics trajectories. This justifies a posteriori the viability of the elastic network approach beyond the expected range of validity of the (quasi)-harmonic free energy approximation (Levitt et al. 1985; Horiuchi and Go 1991; Brooks et al. 1995; Hinsen et al. 2000; Pontiggia et al. 2007).

The model energy F is used to compute, for each protein, the effective matrix, $\tilde{M}$, providing the quadratic potential of mean force acting on the sole degrees of freedom of interest, that is, the positions of the n C_α's marked for alignment. In the following we shall assume that the residues have been re-indexed so that the first n residues (out of a total of N residues) correspond to the marked ones. To illustrate how $\tilde{M}$ is calculated it is useful to divide the M matrix into blocks reflecting the distinction of the degrees of freedom that we wish to retain (the displacement of the first n residues), from the rest:

$$M = \begin{bmatrix} M^a & V \\ V^T & M^b \end{bmatrix} \tag{2}$$

where the superscript T denotes the transpose. The physical interpretation of the blocks is straightforward: M^a corresponds to the interactions among the first n residues themselves, M^b contains the interactions within the remaining N-n residues, and V contains the interactions between the two groups.

The problem of finding $\tilde{M}$ is analogous to the more familiar one of calculating the implicit-solvent force field among amino acids. In that case, M^a would correspond to the bare amino acid interaction in the *explicit*-solvent description, M^b would be the interaction of water molecules, and V the interaction between amino acids and the solvent. Owing to the simple quadratic nature of F in Equation 1, the calculation of the effective energy $\tilde{F}(\delta\vec{x}_1, \delta\vec{x}_2, \ldots, \delta\vec{x}_n)$ governing the effective interaction among the first n residues can be done explicitly (Hinsen et al. 2000; Carnevale et al. 2006, 2007) yielding

$$\tilde{F} = \frac{1}{2}\sum_{i,j=1}^{n}\delta\vec{x}_i\tilde{M}_{ij}\delta\vec{x}_j \equiv \frac{1}{2}\sum_{i,j=1}^{n}\delta\vec{x}_i[M^a_{ij} + \Delta M_{ij}]\delta\vec{x}_j \tag{3}$$

where

$$\Delta M = -V[M^b]^{-1}V^T \tag{4}$$

where $[M^b]^{-1}$ is the pseudo inverse of M^b. It is apparent from the above equation how the interactions mediated by the V matrix complement the bare ones to yield the final ''dressed'' interaction matrix $\tilde{M}$. The lowest-energy modes of the matching residues are finally identified as the eigenvectors associated with the smallest nonzero eigenvalues of $\tilde{M}$.

In the following we shall indicate with $\{\vec{v}_i^{\,\alpha}\}_{i=1,\ldots,n}$ and $\{\vec{w}_i^{\,\alpha}\}_{i=1,\ldots,n}$ the αth low-energy mode of the marked residues for the first and second protein, respectively.

Alignment score

As is customary we shall assume that the 10 lowest-energy modes are sufficient to account for the essential dynamics of the aligned residues (Amadei et al. 1999). Accordingly, the quality of each tentative alignment involving n residues is quantified with the following combined measure of spatial and dynamic consistency:

$$q_n = \sqrt{\max\{0, \frac{1}{10}\sum_{\alpha,\beta=1}^{10}[\sum_{j=1}^{n}\vec{v}_j^{\,\alpha}\cdot\vec{w}_j^{\,\beta}][\sum_{i=1}^{n}\vec{v}_i^{\,\alpha}\cdot\vec{w}_i^{\,\beta}f(d_i)]\}} \tag{5}$$

where d_i is the distance between the C_α positions of the ith aligned residue of the two proteins, and $f(d) = [1 - tgh((d - d_c)/2)]/2$ is a distance weighting factor interpolating the asymptotic values of 0 and 1 for distances, respectively, much larger and smaller than $d_c = 4$ Å. Observe that q_n is bounded between 0 and 1, and its value is one in the case of a perfect correspondence between both low-energy modes and distances of aligned residues (which we recall have been previously optimized by the superposition of the aligned residues).

The measure in Equation 5 does not depend on the choice of the basis of the lowest-energy modes and generalizes the familiar root mean square inner product, $RMSIP = \sqrt{\frac{1}{10}\sum_{\alpha,\beta=1}^{10}\left|\sum_{i=1}^{n}\vec{v}_i^{\,\alpha}\cdot\vec{w}_i^{\,\beta}\right|^2}$ used to measure the consistency of the dynamic spaces of the same protein in two different MD trajectories. The inclusion of the structural modulation, $f(d_i)$, is sought here as we wish to promote not the mere overall dynamic correspondence of matching residues per se but only when these also have a good space proximity. The alignment score is finally defined in terms of Equation 5 as $s_n = q_n - f(n)$ where $f(n) = 0.5115*\exp(-n/336)$ provides the best fit to the approximately exponential trend of the average values of $q(n)$ recorded over the subset of 56 representatives with different topology (see Supplemental material). The subtraction of $f(n)$ ''regularizes'' the raw measure, q_n, allowing a homogeneous comparison of alignments of different length.

Graphical representation of corresponding modes

The score of Equation 5 is invariant upon replacing the orthonormal set of the $\vec{v}$'s (or $\vec{w}$'s) with another one obtained by their suitable linear combination. This property is used to convey in a graphically optimized way the consistency of two sets of low-energy modes. The first optimized basis vector in each set, $\vec{v}'_i = \sum_{j=1,\ldots,10} a_j \vec{v}_j$ and $\vec{w}'_i = \sum_{j=1,\ldots,10} b_j \vec{w}_j$ are found by optimizing the linear weights, a's and b's, so that the scalar product $\vec{v}'_1 \cdot \vec{w}'_1$ is maximum (the unit norm of $\vec{v}'_i$ and $\vec{w}'_i$ is implied). The procedure is iterated to define the remaining vectors of the new basis, which must be orthogonal to those already identified.

Electronic supplemental material

List of representative enzymes; distribution of optimized alignment scores; low-energy modes alignment for hydroxynitrile lyase and haloalkane dehalogenase; generality of the alignment method.

Acknowledgments

We are indebted to P. Carloni, M. Demirel, A. Kohlmeyer, A. Laio, V. Leone, F. Pontiggia, and M. Zampieri for scientific discussions. This work was supported by INFM-Democritos and MIUR grants PRIN-2006025255 and FIRB-2003-RBNE03B8KK.

References

Alexandrov, V., Lehnert, U., Echols, N., Milburn, D., Engelman, D., and Gerstein, M. 2005. Normal modes for predicting protein motions: A comprehensive database assessment and associated web tool. *Protein Sci.* **14:** 633–643.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped blast and psi-blast: A new generation of protein database search programs. *Nucleic Acids Res.* **25:** 3389–3402.

Amadei, A., Ceruso, M.A., and Di Nola, A. 1999. On the convergence of the conformational coordinates basis set obtained by the essential dynamics analysis of proteins' molecular dynamics simulations. *Proteins* **36:** 419–424.

Andreeva, A. and Murzin, A.G. 2006. Evolution of protein fold in the presence of functional constraints. *Curr. Opin. Struct. Biol.* **16:** 399–408.

Atilgan, A.R., Durell, S.R., Jernigan, R.L., Demirel, M.C., Keskin, O., and Bahar, I. 2001. Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophys. J.* **80:** 505–515.

Bahar, I., Atilgan, A.R., and Erman, B. 1997. Direct evaluation of thermal fluctuations in proteins using a single parameter harmonic potential. *Fold. Des.* **2:** 173–181.

Bartlett, G.J., Borkakoti, N., and Thornton, J.M. 2003. Catalysing new reactions during evolution: Economy of residues and mechanism. *J. Mol. Biol.* **331:** 829–860.

Brooks, B.R., Janezic, D., and Karplus, M. 1995. Harmonic analysis of large systems. I. Methodology. *J. Comput. Chem.* **16:** 1522–1542.

Capozzi, F., Luchinat, C., Micheletti, C., and Pontiggia, F. 2007. Essential dynamics of helices provide a functional classification of EF-hand proteins. *J. Proteome Res.* **6:** 4245–4255.

Carnevale, V., Raugei, S., Micheletti, C., and Carloni, P. 2006. Convergent dynamics in the protease enzymatic superfamily. *J. Am. Chem. Soc.* **128:** 9766–9772.

Carnevale, V., Pontiggia, F., and Micheletti, C. 2007. Structural and dynamical alignment of enzymes with partial structural similarity. *J. Phys. Condens. Matter* **19:** 285206. doi: 10.1088/0953-8984/19/28/285206.

Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T., Higgins, D., and Thompson, J. 2003. Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res.* **31:** 3497–3500.

Chothia, C. 1992. One thousand families for the molecular biologist. *Nature* **357:** 543–544.

Delarue, M. and Sanejouand, Y.H. 2002. Simplified normal mode analysis of conformational transitions in DNA-dependent polymerases: The elastic network model. *J. Mol. Biol.* **320:** 1011–1024.

De los Rios, P., Cecconi, F., Pretre, A., Dietler, G., Michielin, O., Piazza, F., and Juanico, B. 2005. Functional dynamics of PDZ binding domains: A normal-mode analysis. *Biophys. J.* **89:** 14–21.

Denton, M. and Marshall, C. 2001. Protein folds: Laws of form revisited. *Nature* **410:** 417. doi: 10.1038/35068645.

Divne, C., Stahlberg, J., Teeri, T., and Jones, T. 1998. High-resolution crystal structures reveal how a cellulose chain is bound in the 50 Å long tunnel of cellobiohydrolase I from *Trichoderma reesei*. *J. Mol. Biol.* **275:** 309–325.

Falke, J.J. 2002. Enzymology: A moving story. *Science* **295:** 1480–1481.

Garcia, A.E. 1992. Large-amplitude nonlinear motion in proteins. *Phys. Rev. Lett.* **68:** 2696–2699.

Gupta, S., Mangel, W., McGrath, W., Perek, J., Lee, D., Takamoto, K., and Chance, M. 2004. DNA binding provides a molecular strap activating the adenovirus proteinase. *Mol. Cell Proteomics* **3:** 950–959.

Halle, B. 2002. Flexibility and packing in proteins. *Proc. Natl. Acad. Sci.* **99:** 1274–1279.

Hinsen, K. 1998. Analysis of domain motions by approximate normal mode calculations. *Proteins* **33:** 417–429.

Hinsen, K., Petrescu, A.J., Dellerue, S., Bellisent-Funel, M.C., and Kneller, G. 2000. Harmonicity in slow protein dynamics. *Chem. Phys.* **261:** 25–37.

Holm, L. and Sander, C. 1996. Mapping the protein universe. *Science* **273:** 595–603.

Holm, L. and Park, J. 2000. Dalilite workbench for protein structure comparison. *Bioinformatics* **16:** 566–567.

Horiuchi, T. and Go, N. 1991. Projection of Monte Carlo and molecular dynamics trajectories onto the normal mode axes: Human lysozyme. *Proteins* **10:** 106–116.

Humphrey, W., Dalke, A., and Schulten, K. 1996. Vmd—Visual molecular dynamics. *J. Mol. Graph.* **14:** 33–38.

Keskin, O., Jernigan, R.L., and Bahar, I. 2000. Proteins with similar architecture exhibit similar large-scale dynamic behavior. *Biophys. J.* **78:** 2093–2106.

Konagurthu, A.S., Whisstock, J.C., Stuckey, P.J., and Lesk, A.M. 2006. MUSTANG: A multiple structural alignment algorithm. *J. Mol. Biol.* **64:** 559–574.

Lange, O.F. and Grubmueller, H. 2006. Generalized correlation for biomolecular dynamics. *Proteins* **62:** 1053–1061.

Lesk, A.M. 2004. *Introduction to protein science: Architecture, function and genomics*. Oxford University Press, Oxford, UK.

Lesk, A. and Fordham, W. 1996. Conservation and variability in the structures of serine proteinases of the chymotrypsin family. *J. Mol. Biol.* **258:** 501–537.

Levitt, M. and Gerstein, M. 1998. A unified statistical framework for sequence comparison and structure comparison. *Proc. Natl. Acad. Sci.* **95:** 5913–5920.

Levitt, M., Sander, C., and Stern, P.S. 1985. Protein normal-mode dynamics: Trypsin inhibitor, crambin, ribonuclease and lysozyme. *J. Mol. Biol.* **181:** 423–447.

Levy, R.M., Srinivasan, A.R., Olson, W.K., and McCammon, J.A. 1984. Quasi-harmonic method for studying very low frequency modes in proteins. *Biopolymers* **23:** 1099–1112.

Lupas, A.N., Ponting, C.P., and Russell, R.B. 2001. On the evolution of protein folds: Are similar motifs in different protein folds the result of convergence, insertion, or relics of an ancient protein world? *J. Struct. Biol.* **134:** 191–203.

Maguid, S., Fernández-Alberti, S., Parisi, G., and Echave, J. 2006. Evolutionary conservation of protein backbone flexibility. *J. Mol. Evol.* **63:** 448–457.

Mayer, K., Earley, M., Gupta, S., Pichumani, K., Regan, L., and Stone, M. 2003. Covariation of backbone motion throughtout a small protein domain. *Nat. Struct. Biol.* **10:** 962–965.

Micheletti, C., Carloni, P., and Maritan, A. 2004. Accurate and efficient description of protein vibrational dynamics: Comparing molecular dynamics and Gaussian models. *Proteins* **55:** 635–645.

Ming, D. and Wall, M.E. 2005. Allostery in a coarse-grained model of protein dynamics. *Phys. Rev. Lett.* **95:** 198103. doi: 10.1103/PhysRevLett.95.198103.

Mol, C., Kuo, C.-F., Thayer, M., Cunningham, R., and Tainer, J. 1995. Structure and function of the multifunctional DNA-repair enzyme exonuclease III. *Nature* **374:** 381–386.

Murzin, A., Brenner, S., Hubbard, T., and Chothia, C. 1995. SCOP: A structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247:** 536–540.

Notredame, C., Higgins, D.G., and Heringa, J. 2000. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* **302:** 205–217.

Pearl, F., Todd, A., Sillitoe, I., Dibley, M., Redfern, O., Lewis, T., Bennett, C., Marsden, R., Grant, A., Lee, D., et al. 2005. The CATH domain structure database and related resources Gene3D and DHS provide comprehensive domain family information for genome analysis. *Nucleic Acids Res.* **33:** D247–D251. doi: 10.1093/nar/gki024.

Pidugu, L., Kapoor, M., Surolia, N., Surolia, A., and Saguna, K. 2004. Structural basis for the variation in triclosan affinity to enoyl reductases. *J. Mol. Biol.* **343:** 147–155.

Pontiggia, F., Colombo, G., Micheletti, C., and Orland, H. 2007. Anharmonicity and self-similarity of the free energy landscape of protein G. *Phys. Rev. Lett.* **98:** 048102.

Porter, C.T., Bartlett, G.J., and Thornton, J.M. 2004. The catalytic site atlas: A resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Res.* **32:** D129–D133.

Rod, T.H., Radkiewicz, J.L., and Brooks, C.L. 2003. Correlated motion and the effect of distal mutations in dihydrofolate reductase. *Proc. Natl. Acad. Sci.* **100:** 6980–6985.

Rose, G., Fleming, P., Banavar, J., and Maritan, A. 2006. A backbone-based theory of protein folding. *Proc. Natl. Acad. Sci.* **103:** 16623–16633.

Roujeinikova, A., Sedelnikova, S., de Boer, G., Stuitje, A., Slabasi, A., Rafferty, J., and Rice, D. 1999. Inhibitor binding studies on enoyl reductase reveal conformational changes related to substrate recognition. *J. Biol. Chem.* **274:** 30811–30817.

Sacquin-Mora, S., Laforet, E., and Lavery, R. 2007. Locating the active sites of enzymes using mechanical properties. *Proteins* **67:** 350–359.

Scheeff, E.D. and Bourne, P.E. 2005. Structural evolution of the protein kinase-like superfamily. *PLoS Comput. Biol.* **1:** e49. doi: 10.1371/journal.pcbi.0010049.

Shatsky, M., Dror, O., Schneidman-Duhovny, D., Nussinov, R., and Wolfson, H. 2004a. BioInfo3D: A suite of tools for structural bioinformatics. *Nucleic Acids Res.* **32:** W503–W507. doi: 10.1093/nar/gkh413.

Shatsky, M., Nussinov, R., and Wolfson, H. 2004b. A method for simultaneous alignment of multiple protein structures. *Proteins* **56:** 143–156.

Sierk, M. and Pearson, W. 2004. Sensitivity and selectivity in protein structure comparison. *Protein Sci.* **13:** 773–785.

Smith, G.R., Sternberg, M.J.E., and Bates, P.A. 2005. The relationship between the flexibility of proteins and their conformational states on forming protein–protein complexes with an application to protein–protein docking. *J. Mol. Biol.* **347:** 1077–1101.

Stockwell, G. and Thornton, J. 2006. Conformational diversity of ligands bound to proteins. *J. Mol. Biol.* **356:** 928–944.

Storey, J. and Tibshirani, R. 2003. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci.* **100:** 9440–9445.

Sulkowska, J., Kloczkowski, A., Sen, T., Cieplak, M., and Jernigan, R. 2007. Predicting the order in which contacts are broken during single molecule protein stretching experiments. *Proteins* **71:** 45–60.

Taylor, W. 2006. Decoy models for protein structure comparison score normalisation. *J. Mol. Biol.* **357:** 676–699.

Tirion, M.M. 1996. Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. *Phys. Rev. Lett.* **77:** 1905–1908.

Xu, H., Aurora, R., Rose, G.D., and White, R.H. 1999. Identifying two ancient enzymes in archaea using predicted secondary structure alignment. *Nat. Struct. Biol.* **6:** 750–754.

Yang, G., Sandalova, T., Lohman, K., Lindqvist, Y., and Rendina, A. 1997. Active site mutants of *Escherichia coli* dethiobiotin synthetase: Effects of mutations on enzyme catalytic and structural properties. *Biochemistry* **36:** 4751–4760.

Zheng, W., Brooks, B., and Thirumalai, D. 2007. Allosteric transitions in the chaperonin GroEL are captured by a dominant normal mode that is most robust to sequence variations. *Biophys. J.* **93:** 2289–2299.